Large Language Models Are Not Robust Multiple Choice Selectors

Year: 2023 | Citations: 347 | Authors: Chujie Zheng, Hao Zhou, Fandong Meng

---

Abstract

Multiple choice questions (MCQs) serve as a common yet important task format in the evaluation of large language models (LLMs). This work shows that modern LLMs are vulnerable to option position changes in MCQs due to their inherent"selection bias", namely, they prefer to select specific option IDs as answers (like"Option A"). Through extensive empirical analyses with 20 LLMs on three benchmarks, we pinpoint that this behavioral bias primarily stems from LLMs' token bias, where the model a priori assigns more probabilistic mass to specific option ID tokens (e.g., A/B/C/D) when predicting answers from the option IDs. To mitigate selection bias, we propose a label-free, inference-time debiasing method, called PriDe, which separates the model's prior bias for option IDs from the overall prediction distribution. PriDe first estimates the prior by permutating option contents on a small number of test samples, and then applies the estimated prior to debias the remaining samples. We demonstrate that it achieves interpretable and transferable debiasing with high computational efficiency. We hope this work can draw broader research attention to the bias and robustness of modern LLMs.